How Useful is Self-Supervised Pretraining for Visual Tasks?

Alejandro Newell, Jia Deng

Princeton University

CVPR 2020

A systemetic evaluation of self-supversied pretraining

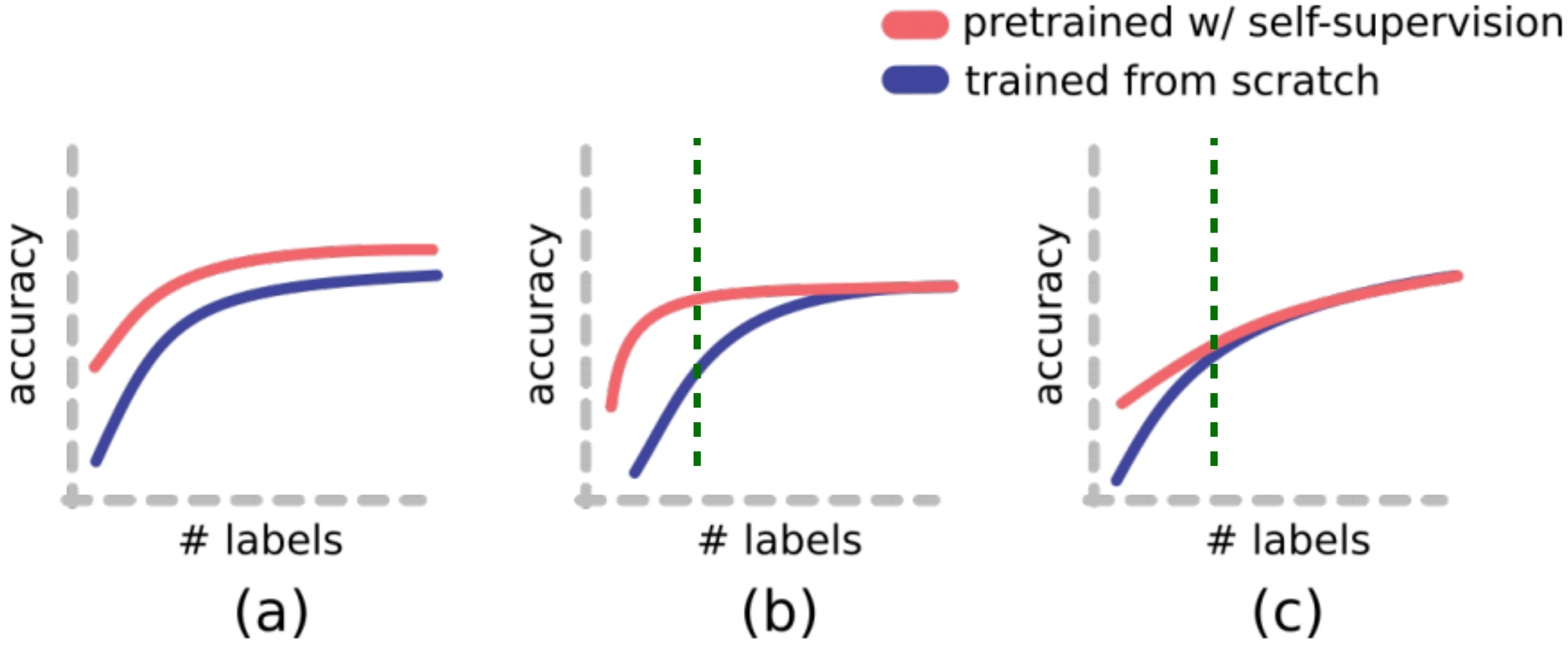
Self-supversied pretraining:
pretraining a network with unlabeled data for later finetuning on a downstream task

A systemetic evaluation of self-supversied pretraining

Self-supversied pretraining:
pretraining a network with unlabeled data for later finetuning on a downstream task

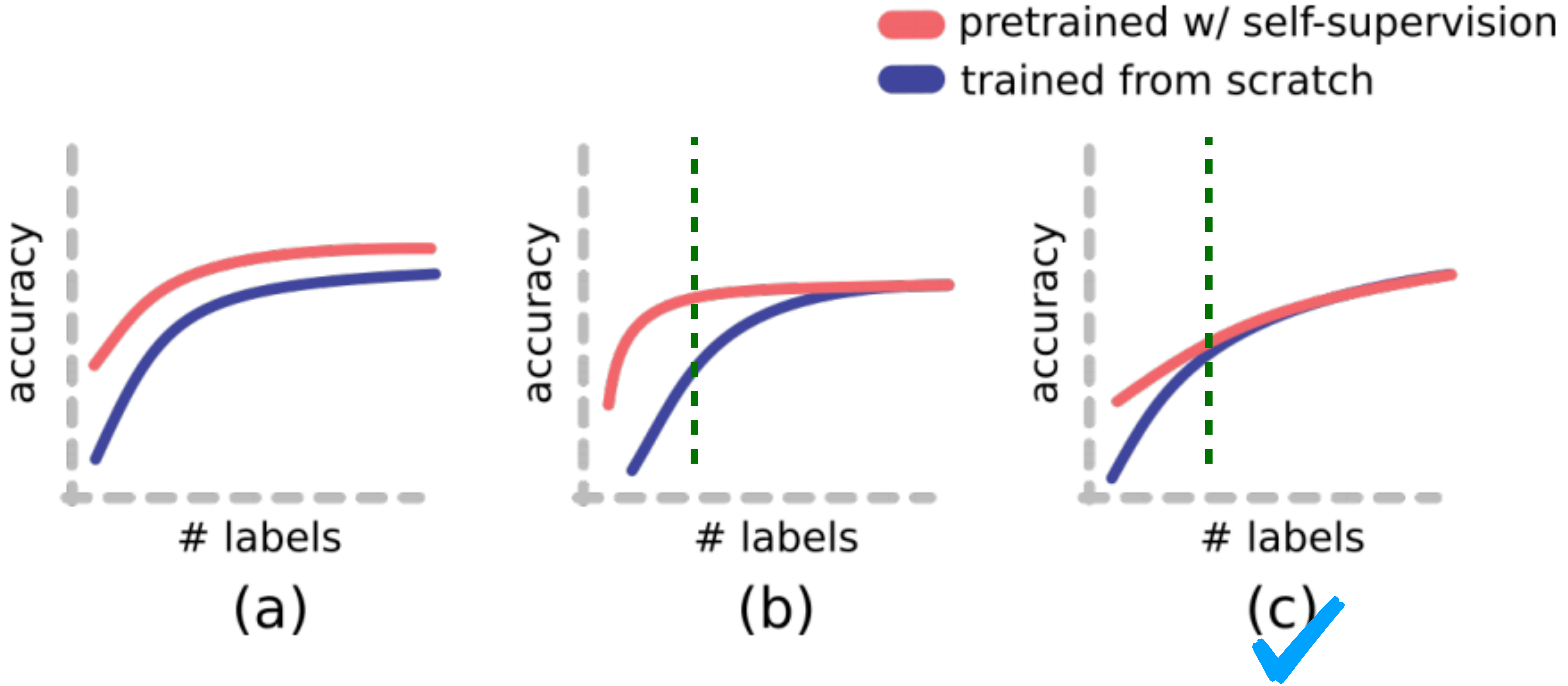
Evaluate its utility by:
comparing a finetuned self-supervised model against a baseline trained from scratch

A systemetic evaluation of self-supversied pretraining



Better accuracy **Reduce Underfitting** **Reduce Overfitting**

A systemetic evaluation of self-supversied pretraining



Better accuracy **Reduce Underfitting** **Reduce Overfitting**

Quantify the utility of self-supervision

$a(n)$ accuracy of a model trained from scratch

$a_{ft}(n)$ accuracy of the finetuned model

$U(n)$ utility at n defined as $\hat{n}/n - 1$

where $a(\hat{n}) = a_{ft}(n)$

This is the ratio of additional labels needed to match the accuracy of the finetuned model

Quantify the utility of self-supervision

Utility vs. Number of labels

Utility vs. Influencing Factors

Quantify the utility of self-supervision

Utility vs. Number of labels

Utility vs. Influencing Factors

Data complexity: *Texture, Color, Viewpoint, Lighting*

Self-supervision algorithm: *VAE, Rotation, CMC, AMDIM*

Model: *ResNet9, ResNet50*

Downstream task: *object classification, object pose estimation, semantic segmentation, and depth estimation (global or dense, semantic or geometric)*

Use synthetic data to control different factors

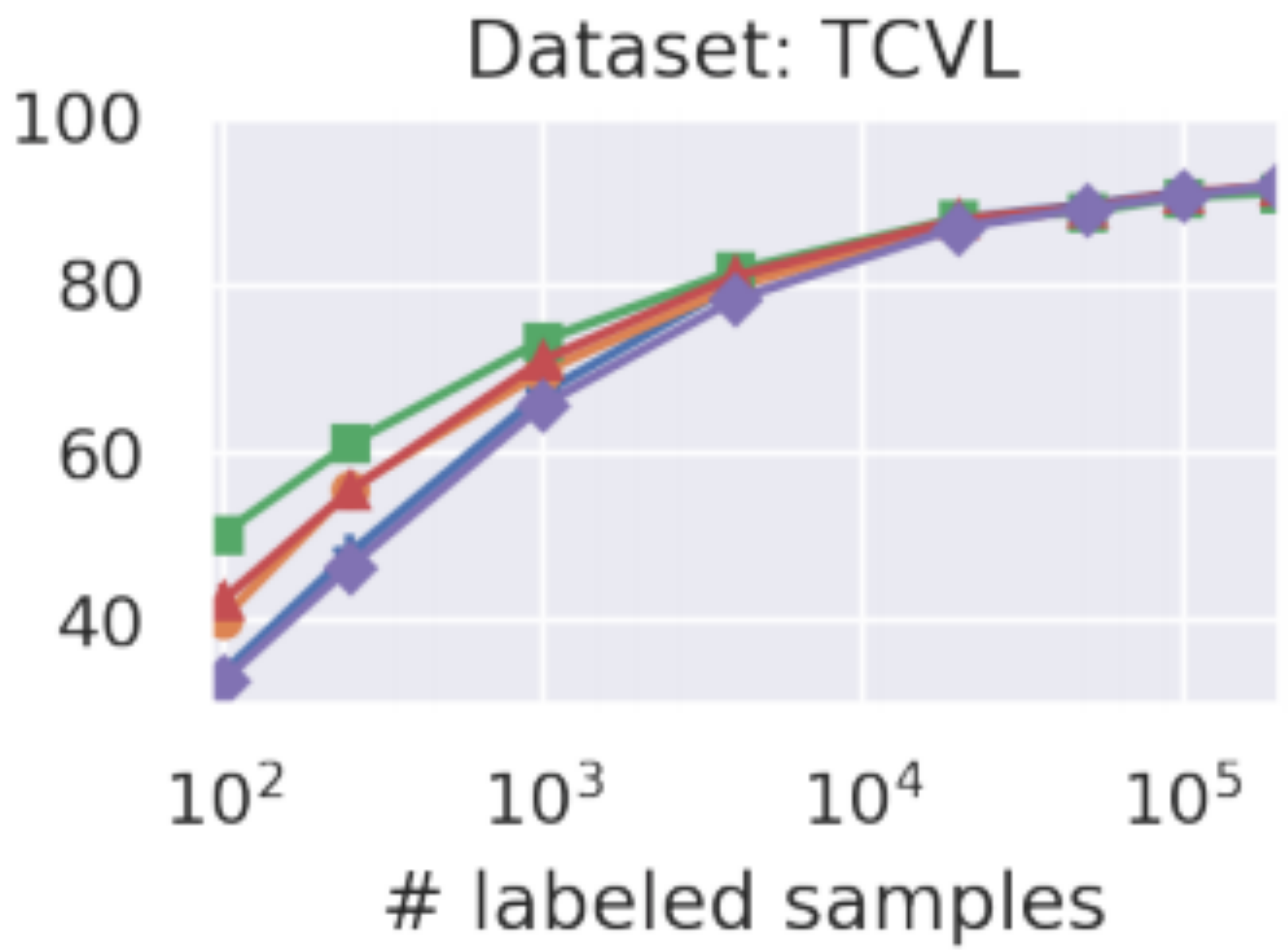
Synthetic data

Figure 2. Example images from four datasets of increasing complexity (from left to right) controlling for viewpoint and texture.

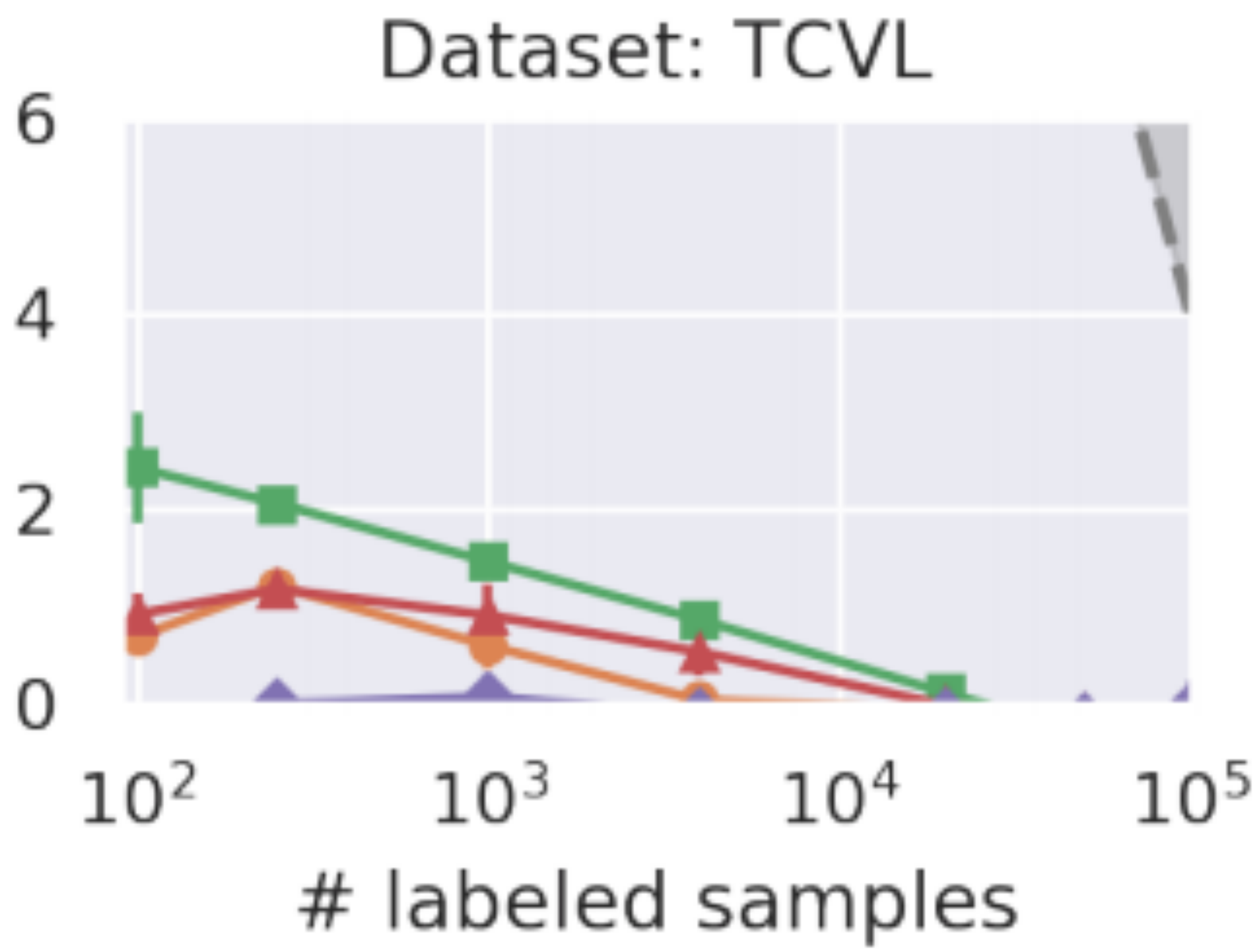
Figure 3. Example images in the multi-object setting as well as the ground truth semantic segmentation and depth.

Finding 1:

self-supervised pretraining methods are useful with a small labeling budget, but utility tends to decrease with ample labels



Finetuned accuracy



Utility

Object classification:
ResNet9

More like a regularization method to reduce overfitting

Finding 2:

Relative performance of methods is not consistent across downstream settings (Evaluation via only classificatino is not sufficient)

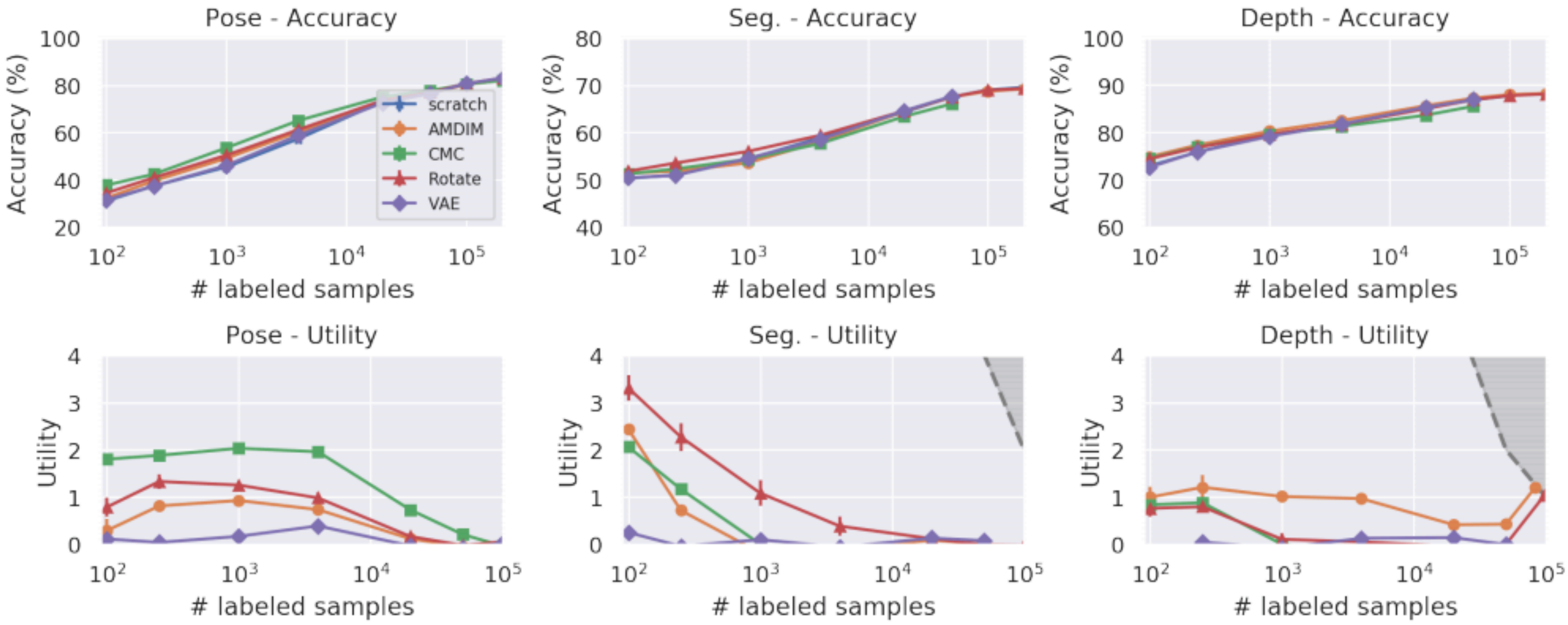


Figure 5. Performance on additional downstream tasks with ResNet9 on the hardest dataset setting (TCVL). The best performing method differs depending on the downstream task suggesting that diverse settings should be considered when comparing self-supervised models.

Finding 3:

More helpful when applied to larger models

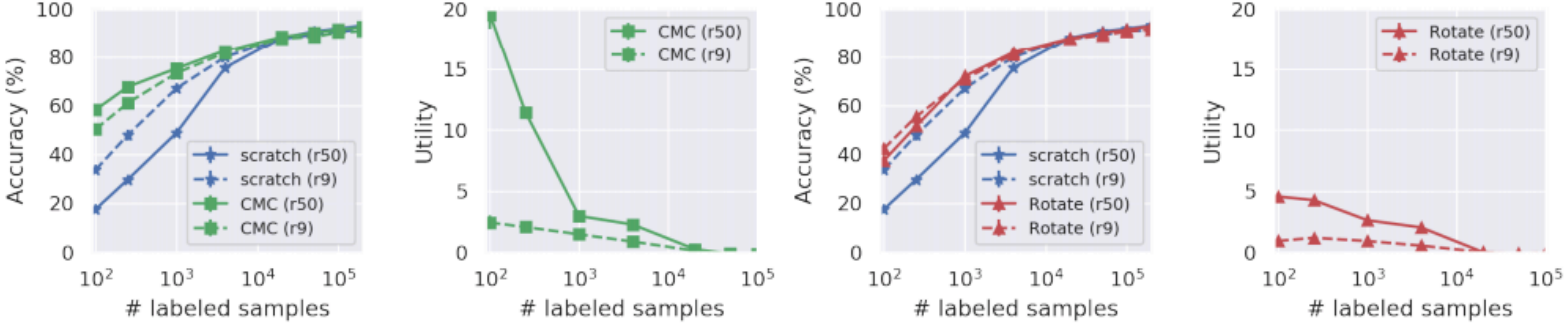
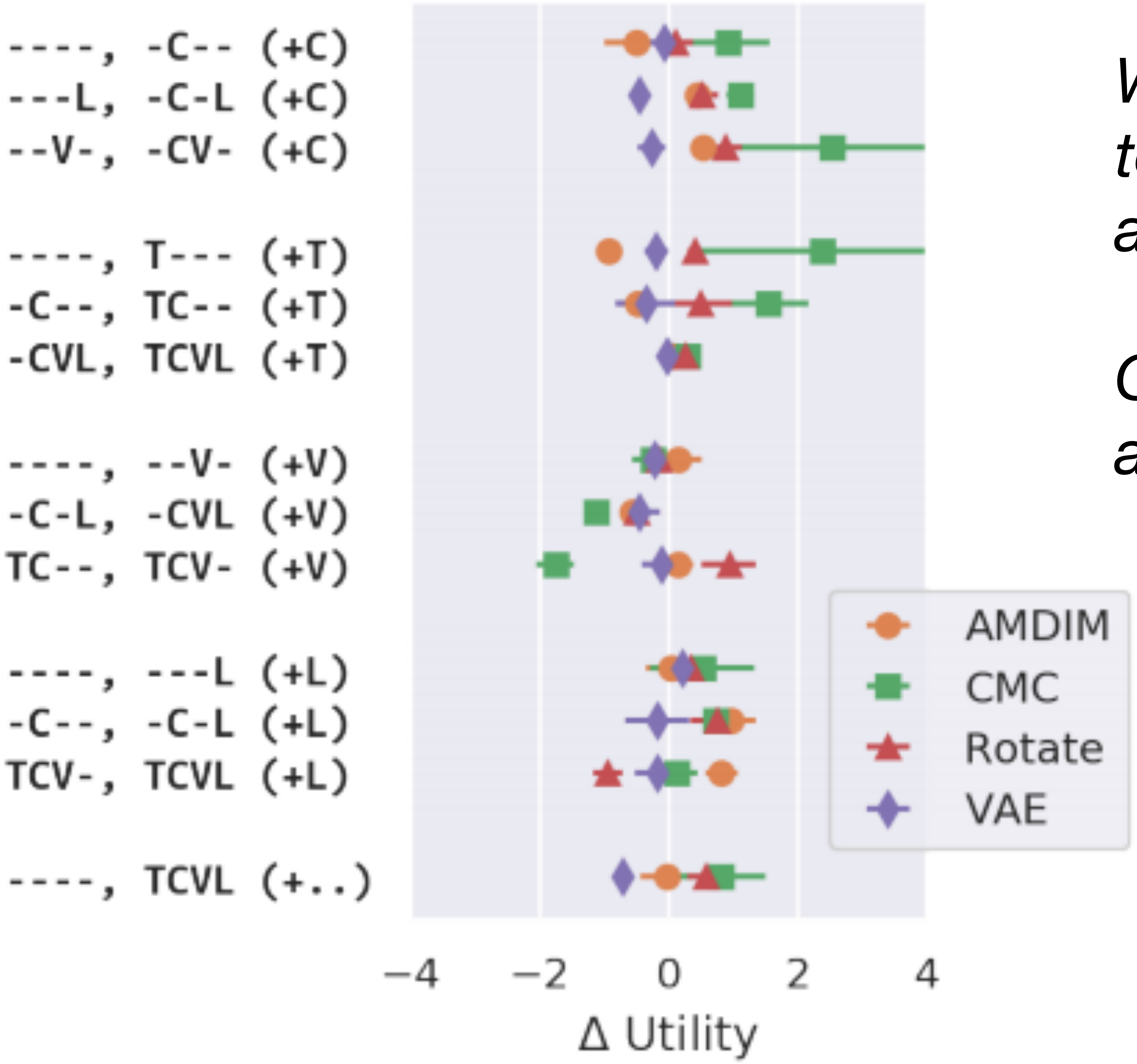


Figure 7. Comparison between ResNet9 and ResNet50 backbones for object classification on TCVL. With few labeled samples the performance of the ResNet50 model is worse when trained from scratch, but when pretrained is better than the pretrained ResNet9 suggesting the importance of pretraining large models when working with less data.

Finding 4:

More helpful when applied to complex data



We observe relatively consistent changes to the utility of a particular algorithm when adjusting a given factor of image variation

Changes to utility for each factor differ across pretraining algorithms

Conclusion

Provide a thorough set of experiments across different downstream tasks and synthetic datasets to measure the utility of pretraining with state-of-the-art self-supervised algorithms

Comments

- Identify flaws of current studies and the limit of self-supervision
- Informative and useful to practitioners